Lab 3 Guide

Relationships Between Variables

This week's lab introduces three ideas: **joint, marginal, and conditional probability**.

- **Joint probability** refers to two events occurring at the **same** time - P(A and B)
- **Marginal probability** refers to the probability of one event occurring **regardless** of the other variable's outcome - P(A)
- **Conditional probability** refers to the probability of one event occurring, **based on a previous event** occurring - P(A|B)

To calculate these values, we first need to calculate the **contingency table** for two variables in a dataset:

pd.crosstab(df.variable1, df.variable2)

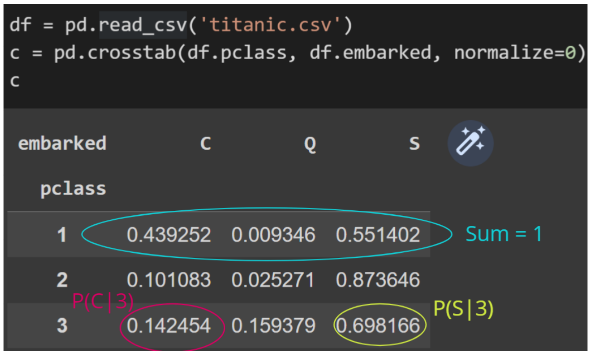
Joint probability can be calculated by simply changing the **normalize parameter to True**:

pd.crosstab(df.variable1, df.variable2, normalize=True)

Marginal probability can be calculated by changing the **margins parameter to True**:

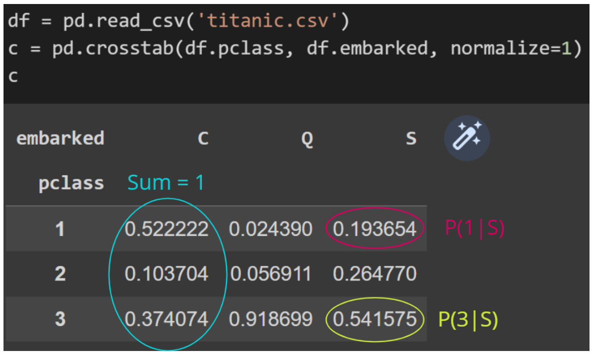
pd.crosstab(df.variable1, df.variable2, normalize=True, margins=True)

Conditional probability can be calculated by passing **0** (to divide by row sum) or **1** (to divide by column sum) into the **normalize parameter**. Below is an example showcasing normalize=0 (embarked given pclass) and normalize=1 (pclass given embarked) on the titanic dataset:



```
df = pd.read_csv('titanic.csv')
c = pd.crosstab(df.pclass, df.embarked, normalize=0)
c
```

embarked pclass	C	Q	S
1	0.439252	0.009346	0.551402
2	0.101083	0.025271	0.873646
3	0.142454	0.159379	0.698166



```
df = pd.read_csv('titanic.csv')
c = pd.crosstab(df.pclass, df.embarked, normalize=1)
c
```

embarked pclass	C	Q	S
1	0.522222	0.024390	0.193654
2	0.103704	0.056911	0.264770
3	0.374074	0.918699	0.541575

The lab also shows how to calculate these probabilities manually using sum().

Exercises 1 and 2 should be straightforward from the example inside the lab. For **exercise 3**, your normalize parameter for crosstab() may be **0 or 1**, depending if the day variable is the horizontal index or vertical index. You have the correct conditional probability when the probabilities for each day **column/row** to add up to **1**.

For each part of **exercise 4**, try to imagine which variable (party size or day) is the **"part"** (numerator) and which variable is the **"whole"** (denominator).

- If the **day** variable is the **"whole"**, then we get:

  2-person parties on saturday / all parties on saturday

- If the **party size** variable is the **"part"**, then we get:

  2-person parties on saturday / 2-person parties on all days